A Search Engine Approach to Estimating Temporal Changes in Gender Orientation of First Names

Brittany N. Smith[1], Mamta Singh[2], and Vetle I. Torvik[3]
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA
[1]bnsmith3@illinois.edu, [2]mamta2.singh@gmail.com, [3]vtorvik@illinois.edu

ABSTRACT

This paper presents an approach for predicting the gender orientation of any given first name over time based on a set of search engine queries with the name prefixed by masculine and feminine *markers* (e.g., "Uncle Taylor"). We hypothesize that these markers can capture the great majority of variability in gender orientation, including temporal changes. To test this hypothesis, we train a logistic regression model, with time-varying marker weights, using marker counts from Bing.com to predict male/female counts for 85,406 names in US Social Security Administration (SSA) data during 1880-2008. The model misclassifies 2.25% of the people in the SSA dataset (slightly worse than the 1.74% pure error rate) and provides accurate predictions for names beyond the SSA. The misclassification rate is higher in recent years (due to increasing name diversity), for general English words (e.g., Will), for names from certain countries (e.g., China), and for rare names. However, the model tends to err on the side of caution by predicting neutral/unknown rather than false positive. As hypothesized, the markers also capture temporal patterns of androgyny. For example, *Daughter* is a stronger female predictor for recent years while *Grandfather* is a stronger male predictor around the turn of the 20th century. The model has been implemented as a web-tool called Genni (available via http://abel.lis.illinois.edu/) that displays the predicted proportion of females vs. males over time for any given name. This should be a valuable resource for those who utilize names in order to discern gender on a large scale, e.g., to study the roles of gender and diversity in scholarly work based on digital libraries and bibliographic databases where the authors' names are listed.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models - *Statistical;* H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Linguistic processing.*

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Gender, androgyny, data mining, temporal prediction, bibliometrics, search engine, textual markers, semantic orientation

JCDL'13, July 22-26, 3013, Indianapolis, Indiana, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2077-1/13/07 …$15.00.

1. INTRODUCTION

Does Taylor refer to a boy or a girl? Answering that type of question can be quite difficult when only given a first name and no other contextual clues. For common, gender-specific names like Michael or Mary humans can answer that question with a high degree of confidence but how can we make a machine coupled with easily accessible digital resources answer that question? Even humans have difficulty with rare or unfamiliar names (e.g., Kuulei or Aagot) because of personal biases such as the number of people that they personally know that happen to be male or female with that name or a similar name. The answer to the question may also vary depending on when the question is posed partly because many androgynous names are increasingly given to girls over time [16]. For example, Taylor was given to more boys than girls prior to 1985, but today, Taylor most often refers to a girl, at least in the USA. In instances like this, can machines do better?

Table 1. Temporal gender markers selected from common first name prefixes in the Google n-gram dataset [31].

	Feminine markers	Masculine markers
1	Mrs	Mr
2	Mother	Father
3	Grandmother	Grandfather
4	Wife	Husband
5	Daughter	Son
6	Aunt	Uncle
7	Sister	Brother
8	Girl	Boy
9	He married	She married

Our motivation stems from analyzing the roles of gender and diversity in the publication and collaboration strategies of authors in large-scale digital libraries and bibliographic databases. This requires an automatic process for confidently assigning gender to each instance of an author name at a certain time point. In this paper, we test the accuracy of a temporal gender prediction model based solely on search engine results as predictors. The model is trained using US Social Security Administration (SSA) data, which includes the frequency with which a name is given to males and females over time. Search result counts are collected from Bing.com queries consisting of temporal and gender-specific markers (see Table 1 for a complete list) concatenated with the name under consideration. Using multiple markers should not only help make predictions less sensitive to noisy search engine query results but should also help capture more nuanced changes in individual names over time. Relying solely on web searches as predictors has several advantages: web searching is easy to automate and thus scales well, the data is public and not domain or genre specific (i.e., it covers personal pages, industry sites, news, social media, blogs, genealogy documents, etc. not just one of these), nor is the data region specific (i.e., not just English names).

It should be noted that a last name can also be useful for gender attribution in some cultures (e.g., Slavic last names ending in -ova or Icelandic names ending in -dottir are given to females). We chose not to train the model using last names because this phenomenon is much less common globally and is rare in the USA specifically. However, one can still use the trained model to make predictions using a last name (or a combination of first and last name) as input.

2. RELATED WORK

Our approach builds on several different areas of work including work that has looked at the manual assignment of gender using names, the automated assignment of gender, and work on semantic orientation.

2.1 Manual Gender Assignment

Several recent studies have looked at gender differences in disparate areas of academia and industry. Though some studies used data from large centers, e.g., the National Center for Education Statistics [12], which one would assume has rather accurate and complete data including individuals' gender, other studies used more imprecise methods to discern gender. These methods included using publications and the author's first name to guess the gender, searching for author information online, or by looking at pronoun usage in biographical notes and/or author bylines, e.g., [6, 21, 22]. While this essentially manual process yields fairly accurate information, it takes a great deal of time to cultivate the data, and if no supporting data can be found to help ascertain gender, sometimes names are simply excluded from the analysis.

2.2 Automated Gender Assignment

In addition to manual assignment, some models have been created to infer gender using automated means. Otterbacher [18] explored the contribution of style, content, and metadata from movie reviews in inferring the gender of movie reviewers on the Internet Movie Database site. Cucerzan and Yarowsky [9] used nouns with known attributed genders to create a bootstrapping process that determines the grammatical gender of other nouns in five different languages, and Bergsma, Lin, and Goebel [4] created a model that uses contextual, morphological, and categorical gender features to classify nouns in documents. The Never-Ending Language Learner developed by Carlson et al. [7] also uses information about noun phrases and other language and text patterns to guess gender in cases where appropriate. We too utilize markers to help infer gender, but we use the help of a search engine.

In two recent papers, Zheleva and Getoor [33] and Tang et al. [23] use user profile data from social network sites like Facebook in order to infer information about individuals including gender. Zheleva and Getoor [33] create eight different models that use varying amounts of an individual's profile data and at times even the individual's friends' data in order to approximate things like gender and political view. Our models use less information and information that is publicly available to achieve a similar goal.

Though Tang et al. [23] also utilize social network data, data from Facebook, they take an approach similar to ours in that they use publicly available data and words therein (e.g., relationship status and who the person is “interested in” or “looking for”) to build models that attempt to discern an individual's gender. Our models differ in that we use less personal, easily attainable data by using general search engine results augmented simply by common gender-specific words and phrases. Our models are also different in that we can look at the way that the name was assigned to people of different genders over time while Tang et al. [23] only evaluate the present usage of names.

In addition to the aforementioned models, there are some websites that provide information on the gender of names, and perhaps not surprisingly, they are mostly proposed as sites to help with the naming of children. Two such tools include NameVoyager [30] and its accompanying site Baby Name Wizard [3] and Baby Name Guesser [2]. NameVoyager is a tool that was initially built to display SSA data from 1900 to 2001 in a way that promotes discovery and reflection. It has since grown to include more recent and detailed information. Baby Name Guesser is a tool that uses Google information to calculate name statistics and includes a gender predictor [2]. It is not made public how Google information is used, but it is perhaps the most similar to our strategy.

2.3 Semantic Orientation

Using known characteristics of some words to determine the orientation of others is not a new idea. The idea of semantic orientation is discussed by Turney and Littman [27] who infer the orientation of a word from its statistical association with other words. Efron [11] used the work of Turney and Littman [27] to build a model that uses cocitation information to estimate the political orientation of web documents. Though we do not use pointwise mutual information or latent semantic analysis measures, the idea of using one word with a known association to infer the association of another is what we utilize when creating Bing API web search queries to predict gender for given names.

3. TRAINING A GENDER PREDICTION MODEL

This section describes how we built a prediction model using US SSA data on gender counts as the target and results of web search queries submitted to the Bing API [5] as predictors.

3.1 Data

The US SSA provides data consisting of the frequency of names and gender on Social Security card applications for each year going back to 1880. To protect privacy, names are only included in the yearly SSA data if the name was given to at least 5 people during that year [20]. From those sets, we included only names that appeared during the 1880 to 2008 time span. This left 85,406 distinct names, which varied greatly in popularity with some names assigned to individuals only 5 times during that time span, e.g., Zaccariah and Ahmeer, to names given to over 5,000,000 individuals, e.g., James and John.

For each name in the dataset, 258 dimensions of SSA data was recorded: 1) the number of females for each of the 129 years during the 1880 to 2008 time span (e.g., there are 2,514 females with the name Alicia in 2004 and 2,333 in 2005) and 2) the number of males for each of the 129 years during the 1880 to 2008 time span. Also, for each name, the results of 72 Bing API web searches were recorded as follows: the estimated total number web pages that contain the phrase made up of the name prefixed by one of 18 markers (see Table 1), for each of 4 different Bing offset parameters (0, 10, 100, and 1000). Note that Bing gives different estimates depending on the offset setting. Table 2 illustrates this effect for the name Genni where, for example, the phrase “Mother Genni” resulted in 38, 7, 7, and 7 results for each of these respective offsets. Query variability is due to the way the Bing API is designed to take into account certain factors like query popularity and the set of results requested (i.e., asking for the 11-20 results to be returned as opposed to requesting the first ten results) [1]. By recording results from different Bing offsets, this parameter setting can be optimized for the predictive model. Using 18 markers also helps deal with query variability. Weights are distributed across the markers which helps reduce the potential effect due to noise in any one marker count. Each search was formulated to ensure that no query correction occurred and the exact phrase was found excepting punctuation marks. For example, +"Mother Jean," will include "mother. Jean" but not "mother of Jean."

Table 2. Marker counts for the name "Genni".

Offset	0	10	100	1000
Mrs vs. Mr	88 vs 27	88 vs 5	16 vs 5	16 vs 5
Mother vs. Father	38 vs 5	7 vs 1	7 vs 1	7 vs 1
Grandmother vs. Grandfather	5 vs 0	1 vs 0	1 vs 0	1 vs 0
Wife vs. Husband	94 vs 11	94 vs 1	16 vs 1	16 vs 1
Daughter vs. Son	44 vs 33	8 vs 3	8 vs 3	8 vs 3
Aunt vs. Uncle	116 vs 0	116 vs 0	17 vs 0	17 vs 0
Sister vs. Brother	44 vs 5	8 vs 1	8 vs 1	8 vs 1
Girl vs. Boy	66 vs 11	66 vs 2	12 vs 2	12 vs 2
He married vs. She married	0 vs 0	0 vs 0	0 vs 0	0 vs 0

3.2 Logistic Regression Modeling

Given a name characterized by an 18-dimensional vector of marker values x, the purpose of the model is to estimate the proportion of individuals that are female for the given x and $year$: $p(Female|x; year)$. The i-th element of the vector x_i represents a gender predictor and is encoded by transformed Bing® results for each of the following markers: x_1 = Mrs, x_2 = Mr, x_3 = Mother, x_4= Father, x_5 = Grandmother, x_6 = Grandfather, x_7 = Wife, x_8 = Husband, x_9 = Daughter, x_{10} = Son, x_{11} = Aunt, x_{12} = Uncle, x_{13}= Sister, x_{14} = Brother, x_{15} = Girl, x_{16} = Boy, x_{17} = He married, x_{18} = She married. Each marker value was defined by a simple transformation of the corresponding Bing count as follows:

$$x_i = adjust(\log(bing\_count + \delta))$$

A δ correction was applied to avoid undefined values for the natural log, and two different values $\delta \in (1,5)$ were tested for their effect on the predictions. When plotting the results of the transformation, we noticed fixed artificial gaps, presumably due to Bing's estimation procedure (see Figure 1 for an example). The gaps identified for each offset value were as follows: gap = 2.7-3.6 for 0 offset; gap = 1-1.8 and 2.7-3.6 for 10 offset; gap = 2-2.8 and 3.6-4.05 for 100 offset; no gaps for 1000 offset. The *adjust()* function creates a continuous distribution by collapsing the gaps as follows: subtract the gap size for points above the gap and shift the points inside the gap to the lower bound.

A weighted combination of the transformed attribute values, using a logistic regression model, is used to estimate the probability. Temporal changes in the weights assigned to the markers were captured in two different ways. The "full" model creates a quadratic fit for each marker weight across the entire 129 years, while a "sliding-window" model fits separate linear models to each 9-year window across the 129 years. The "sliding-window" model should capture more local changes over time, while the "full" model smoothes out the local changes and only captures global trends over time. The logistic regression models can be expressed in the following generic form:

$$\begin{aligned} logit(p) = & \sum_{i=0}^{18}(\alpha_i + \beta_i y + \gamma_i y^2)x_i \\ & + \sum_{i=1}^{18}(\alpha_{i+18} + \beta_{i+18} y + \gamma_{i+18} y^2)(x_i = 0) \\ & + \sum_{i=1}^{18}(\alpha_{i+36} + \beta_{i+36} y + \gamma_{i+36} y^2)(0 < x_i \le c) \end{aligned}$$

Each x_i refers to a transformed attribute value and each $\alpha_i + \beta_i y + \gamma_i y^2$ corresponds to the weight assigned to that attribute for a given year = $y \in (1880, 1881, \cdots, 2008)$ for the "full" model. Note that we define $x_0 = 1$ to capture the intercepts. The "sliding-window" model excludes the quadratic year terms (i.e., $\gamma_i = 0$). Two sets of indicator attributes capture effects due to low counts: whether x_i is 0 or not, and whether x_i is a small positive value $\le c$. The values for c were set empirically: $c = 5$ for offset = 0, and 1 for all other offsets. The "full" model is trained on all observed names across all 129 years whereas the "sliding-window" models consist of a series of submodels each of which is trained using observations for a given year +/- a window of 4 years: $y \in (1880 - 1888,\ 1889 - 1897,\ \cdots, 2000 - 2008)$. The reason for fitting the less restrictive "sliding-window" model is to test whether the "full" model captures the full extent of temporal change in weights. Overall, the models were trained with a combination of different parameter settings: $\delta \in (1,5)$, $offset \in (0,10,100,1000)$, and with and without the indicator attributes. This gives sixteen "full" models and the sixteen "sliding-window" models from which to select.

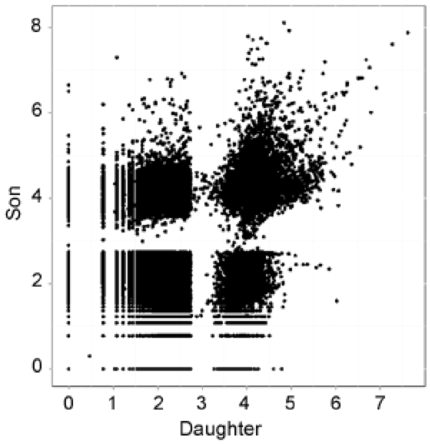

Figure 1. Bing result counts (δ = 1, offset = 0) for all names in the SSA dataset. Note the sharp separation of points reflecting Bing's step-estimation for values above a fixed cut-off.

4. RESULTS AND DISCUSSION

4.1 Overall misclassification rates and model selection

Figure 2 shows the proportion of individuals in the SSA dataset misclassified by some of the models over time and compared against the pure (within name*year and within name; labeled SSA w/year and SSA w/o year, respectively) error rates under two different classification scenarios. The first scenario requires classification of all individuals as either male or female (i.e., probability > 50% implies female, otherwise male) which yields higher error rates than the second scenario which permits unknown or neutral classifications (where 10% < probability < 90%) that are excluded from error calculations. The second scenario captures errors on highly confident predictions only.

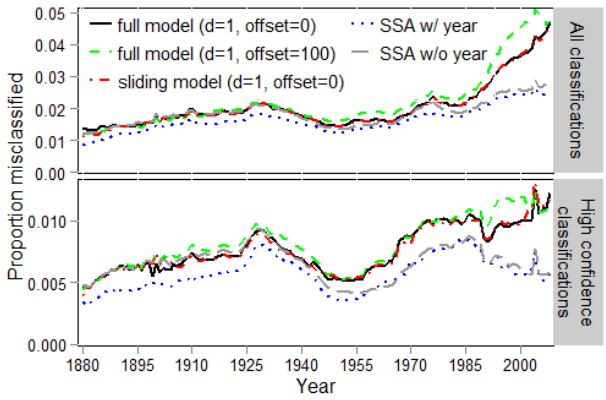

Figure 2. The proportion of individuals in the SSA dataset misclassified over time under two different scenarios: *All classifications* (female if p > 0.5; male otherwise) and *High confidence classifications* only (female if p > 0.9, male if p < 0.1, unknown otherwise). *SSA w/ year* corresponds to the pure (within name*year) error and represents a lower bound for any model. *SSA w/o year* refers to within-name error; and its difference with *SSA w/year* reflects total error due to time.

The models with the indicators produced more consistent and lower error rates than the models without indicators (not shown), and there were no effective differences in error curves between $\delta = 5$ vs. $\delta = 1$ (not shown). The error curves for the “full” and “sliding-window” models are very similar, which implies that the sliding-window approach to capture local time variations does not contribute to the overall performance. The offset matters though. The “full” model with offset of 0 performs better than all other offsets (“full” model with offset 100 is shown in Figure 2), particularly in more recent years. Out of the thirty-two models, the indicator “full” and the indicator “sliding-window” model with offset of 0 and $\delta = 1$ were selected, and the remainder of the paper will utilize either one or both of these models. They will be referred to as the “full” and “sliding” model, respectively.

The “full” model misclassifies 2.25% of the people in the SSA dataset, which is slightly higher than the 1.74% pure (within name*year) error rate; when excluding low confidence predictions the model has an SSA misclassification rate of 0.82% vs. 0.59% pure error. Figure 2 shows that the misclassification rates range from 1% to 4.5% over the entire 129 years, while the high confidence rates range from 0.5% to 1%. Overall the misclassification rates and pure errors rates tend to increase over time, likely due to increasing diversity of names and the fact that more names are being given to individuals of both genders than in the past. The downturn during ~1930-1950 probably reflects, in part, lower immigration rates during the Depression and World War II.

It should also be noted that the SSA data is not 100% reliable as a gold standard for two reasons. First, the SSA data is likely to contain human recording errors e.g., from the transcription of newly minted parents hand-written notes on paper forms. It would not be surprising if this occurred at a rate of ~1%, e.g., for older records, but is more likely to occur at a scale 0.1%. Second, the model predictions are based on mentions of the Bing-indexed names across the world-wide web and these mentions represent a sample of a population that is much broader than the one captured in the SSA dataset. As a result, the model can provide better worldwide estimates than the SSA data, e.g., for names like Andrea which is dominated by females in the US but dominated by males in Italy. We will further discuss this effect in subsequent sections.

4.2 Characterizing the model

Figure 3a shows the weights of the phrase markers (and their indicator attributes in Figures 3b and 3c) as they change over time in the two selected models. Overall, the “full” model weights change smoothly over time capturing general trends, while the “sliding” model captures short term changes as shown by fluctuations and occasional spikes. This is because the “full” model takes into account data from all 129 years using a quadratic approximation while the “sliding” model is based on separate linear approximations for each possible 9-year window.

The marker weights vary in systematic and meaningful ways, both in direction and absolute values. For example, *He married* and *She married* were the most heavily weighted positive (i.e., female) and negative (i.e., male) markers, respectively, during the earliest years but have gradually lost about 50% of their weights during the time period covered. Married references often point to genealogy and family tree descriptions going back 100 or more years. The weights for *Grandmother* and *Grandfather* are much smaller but follow a similar pattern, except that they switch direction (*Grandmother* becomes male; *Grandfather* becomes female) in the 1930s. *Daughter* and *Son* (and *Girl* and *Boy*) follow a similar pattern, except time-reversed, while *Sister* and *Brother* peak in the middle of the time period, as does *Wife* and *Husband*. It makes perfect sense that mentions of grandparents, the “oldest” markers, peak at the earliest time, while children, the “youngest” markers, peak today, and other family members peak in between.

Interestingly, the gender-opposite paired curves tend to vertically mirror one another, except for *Father* and *Mother*. We expected this mirroring to occur across the zero-line with the feminine indicator being positive and the masculine indicator being negative. Many of the pairings followed this pattern but some did not. *Boy*, for example, has a positive weight until the late 1960s.

It is not clear why this is the case but it should not matter much. It is important to note that the model takes into account all markers simultaneously, and the anomalous weights make up a fraction of the total weight. Also, some of the well-behaved markers, *He married* and *She married* and *Aunt* and *Uncle* for example, also tended to result in lower, more accurate Bing counts than the other phrase markers such as *Boy* and *Girl*. As these counts reach thousands or more, Bing starts returning approximations instead of actual result counts, which was the case with most of the queries initiated using the more popular phrase markers like *Mr* and *Father*. It may be the case that these approximations and the large numbers themselves affected the weight sizes and direction.

The weights associated with the zero and low count indicators show similar patterns across the paired phrase markers (Figures 3b and 3c). The indicators were used to relax the continuity of the models near the zero and small Bing counts. We expect the weights to have opposite signs from the transformed Bing result count weights. If a *Daughter* query returns few or no results, we would expect the name to be more likely masculine, which is why in Figure 3b and Figure 3c we mostly see that the masculine indicators have positive weights and the feminine indicators have negative weights in all cases except the *He married* and *She married* pairing. This probably just reflects the fact that they were, by far, the most likely to return zero or low counts.

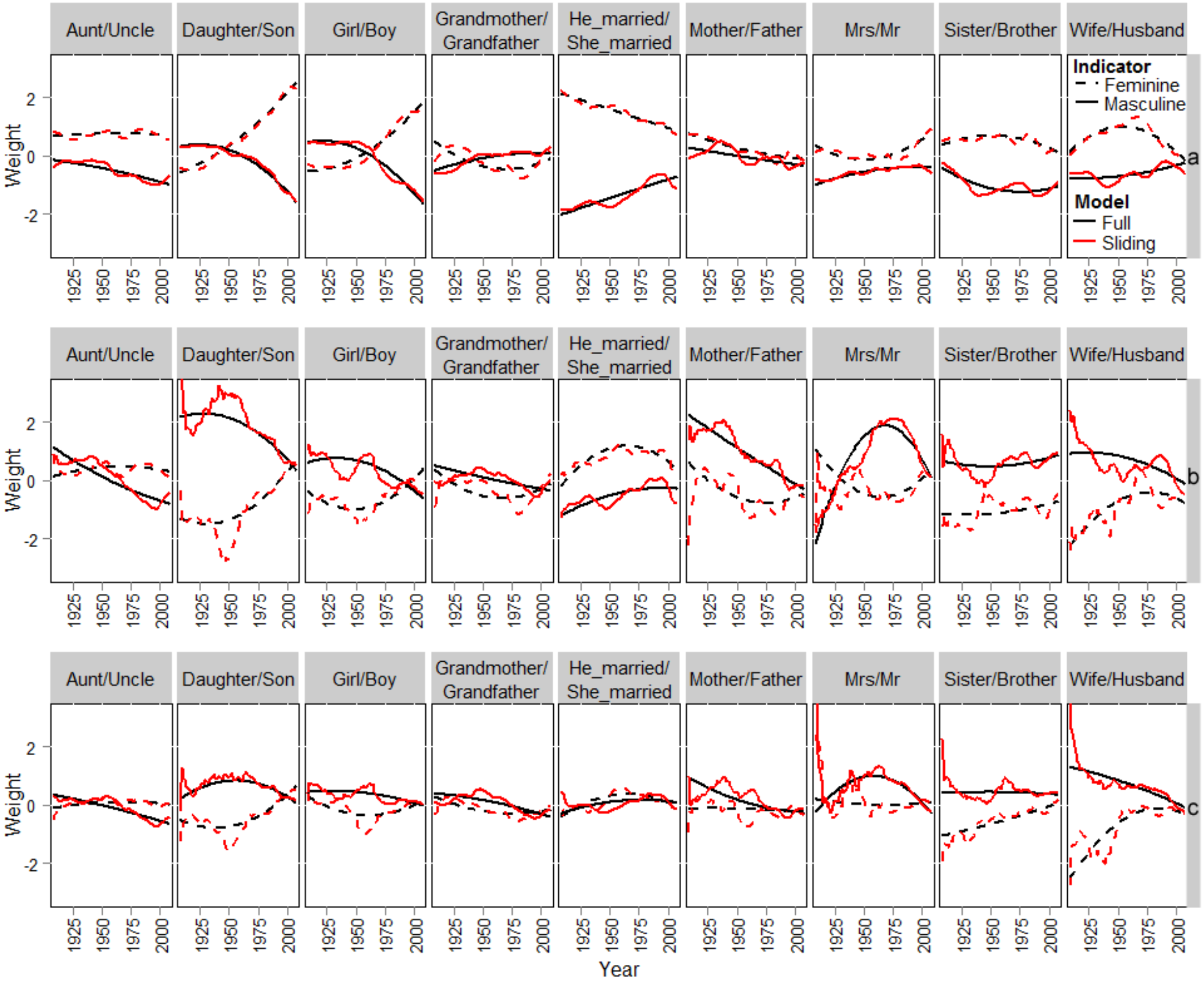

Figure 3. Time-varying weights of the markers: a) transformed Bing counts, b) zero count indicators, c) low count indicators.

4.3 Model performance on select names

We first tested cases where the models are expected to perform well: four popular male names (William, Andrew, Thomas, and Joseph) and four popular female names (Katherine, Julia, Mary, and Catherine). As shown in Figure 4, the predictions vary little and follow the "true" SSA proportions. Also, note that the "sliding" model mimics some name-specific SSA spikes a bit too closely, which suggests that it is over-trained. Keep in mind that the error rates are very low, at a scale that is probably approaching the human clerical error rate.

Second, the models were tested on cases where we expect noisy predictions: ten androgynous names (Alex, Dana, Jackie, Jessie, Jordan, Leslie, Marion, Pat, Terry, and Willie). As Figure 5 shows, these names either switch genders over time (as with Leslie, Dana, and to some extent Willie), remain consistently androgynous (as with Pat), or they change moderately but remain largely attributed to one gender (as with Alex, Jackie, Jessie, Jordan, Marion, and Terry). Overall the models do surprisingly well in identifying these as androgynous names and capturing the direction of gender-orientation changes over time. As expected, the models smooth out highly non-linear changes such as the step-like pattern observed for Leslie. Note also that some predictions differ significantly from the SSA, partly reflecting the global nature of the model (vs. the US-centric SSA data). Both of these effects are beneficial for generalization purposes.

Third, we used a set of author names on computational linguistics papers labeled as male or female using a combination of database lookup and manual annotations[29]. Within this set, the full model (female probability averaged over the years 1955, 1965, and 1975) agreed with 91.9% of 2,755 female labels, and 95.6% of 6,862 male labels. The model strongly disagreed (female prob. < 0.05) with 16 of their female labels (Deryle, Gerik, Helmar, Steffan, Craige, Chikara, Ransom, Chul, Muath, Stephane, Ryosuke, Yuya, Elvan; some occurred multiple times) most of which actually refer to males. The model strongly disagreed (female prob. > 0.95) with 18 of their male labels (Itziar, Sina, Raymonde, Einav, Vivi, Xiaohong, Michele, Sanja, Korin, Ales, and Dorry; some occurred multiple times) several of which actually refer to females.

4.4 Sources of errors

The models are trained on popular US names and will therefore tend to agree when the SSA indicates that a name is strongly female or male. However, there are numerous cases when the models differ from the SSA data. This is partly due to the fact that the predictive markers are based on a global resource (the web as captured by Bing). As Figure 5 shows, the model predictions systematically differ from the SSA proportions for Marion, which is largely given to females in France and other parts of Europe but is given to both males and females in the US. Andrea, Michele, Ira, Jan, and Joan are other examples that have different genderorientations in the US vs. other places in the world. To study these types of variations systematically, we created nine different subsets of names. Figure 6 shows the overall proportion misclassified for each set using the 50% cutoff. Below we discuss these sets and the sources of errors therein.

4.4.1 Non-US Names

We first compiled a list of common non-US first names with a high co-occurrence rate with non-US affiliations (relative to US co-occurrence) in PubMed. The model performs surprisingly well for these names (see grey dashed line in Figure 6a). The error rates are lower than for all SSA names (solid black line) but worse than for common US names (red dashed line). Even though these names are more popular in other countries, many are also present in the US e.g., Isabella and Zoe. Tightly gendered names in some non-US countries (e.g., due to national naming laws) is likely to lead to greatly skewed Bing results with high counts for masculine markers and low to no counts for feminine markers or vice versa.

Second, we assess model performance on subsets of names that are common in countries with recent immigration to US: 300 popular Mexican names [17] and 620 Asian names [19], including but not limited to Chinese, Japanese, Indian, Korean, and Vietnamese names. Figure 6a shows the misclassification rates using SSA numbers as ground truth. Some spikes in the Asian names and Mexican names reflect individual names: Jo (1931), Kim (1954), and Jaime (1976). Jaime's spike in 1976 represents a drastic increase in females named Jaime, and coincides with the introduction of the popular US TV series Bionic Woman and the lead character Jaime Sommers. Jo and Kim are popular diminutives in the US. Setting aside these three names gives us a better perspective of the collective effect of Asian and Mexican names. Error rates for Mexican names are about the same as error rates averaged across all SSA names, while Asian names have higher error rates, particularly since ~1970 and onwards. There are four major reasons for higher error rates amongst names in this group: 1) they are typically short names; 2) transliterations from the original language into English is many-to-many (e.g., the gender-specific aspects of Asian characters or symbols are lost in translation to ASCII); 3) they are typically more androgynous than English names; and 4) they are often used as both first and last names. So, searching for +"Brother Kaoru", for example, may yield results where Kaoru is one's surname and not one's first name.

In some countries, names are often made up of compound combinations of male, female, and androgynous names such as Jose Maria (male in Spain) vs. Jose (male) and Maria (female). The model tends to do well in these cases, as long as the names are relatively common.

4.4.2 Surnames, famous people, and common words

Carlson and Cooper are examples of surnames that are also used as first names, and they are both typically given to males (see Figure 7). As discussed earlier, androgynous names are typically given to more females over time, yet the model predictions indicate that Cooper defies this trend by moving from being somewhat feminine to almost completely masculine in recent years. Carlson, on the other hand, does trend towards becoming more feminine as time progresses. Generally, surnames will give high counts for both the *Mr* and *Mrs* markers which forces the predictions towards neutral.

Unlike names that serve as both first names and surnames, one would expect the names of famous people to be predicted strongly feminine or masculine. Paris (the city and the famous socialite Paris Hilton), Theresa (Mother Theresa), and Guadalupe (the Virgin of Guadalupe, the Roman Catholic icon of the Virgin Mary) are three such names. Both Guadalupe and Theresa are both strongly feminine, as expected, while Paris ends up feminine after climbing there steadily over almost 60 years (see Figure 7). This might be another case where a single famous person has shifted a global trend (also see previous discussion of Jaime).

Given a recent trend to choose nontraditional names for babies, we were also curious how well the model performs for names that are also common English words (including the markers themselves). These names were downloaded from a British baby name website [32] to which we added seven colors in the SSA dataset. As Figure 6b shows, misclassification rates are high for word names, particularly in recent years because of an increase in name creativity. Figure 8 shows predictions of four word names typically attributed to males (i.e., Cash, Given, Hero, and Will) and four word names typically attributed to females (i.e., April, Lily, May, and Summer). While April, Lily, and Summer are given to a greater proportion of females over time, May becomes more masculine over time. The male names, with the exception of Given, remain mostly attributed to males throughout the entire period. Verb names like May and Will are expected to be problematic because many of the markers are nouns, and so search queries become commonly used noun-verb phrases. As with surnames, the resulting word-name predictions tend to err towards neutral rather than false positive male or female.

The name Given is an interesting example in this set of names. As expected, the model predicts more feminine over time but the change occurred earlier than expected (in the 1950s). This is due to the way search engines process punctuations when indexing documents, and how specific phrasing is used across domains. For example, most results from the queries *He married Given* and *She married Given* without punctuation were from obituaries, newspapers, alumni newsletters, genealogical records, and the like. Results from *He married Given* with punctuation interspersed, however, were common in dating advice sites, such as: "but I doubt he's married, given the information..." and "...admitted he's never been tested before he married. Given this information about both our pasts..." The marker *He married* is feminine-oriented in the model. So, despite the baby name site suggesting Given is a masculine name, the model moves towards female prediction because these dating advice sites amplify the *He married* marker relative to the *She married* marker. These findings do not necessarily imply that one should attempt to exclude these types of search results generally, nor is it to say that the *He married* and *She married* markers should not be utilized. This does illustrate, however, that some common words may create unanticipated phrasing that might yield a strong signal in the wrong direction.

Figure 9 shows the predictions for colors used as names in the SSA dataset. The predictions indicate that colors were used only for boys 100 years ago. Since then, Blue, Brown, Green, Orange, and Red remained male, while Pink, White, and Cyan follow the pattern noted by Lieberson, Dumais, and Baumann [16] and become increasingly feminine over time. Some of these colors are strongly attributed to a certain gender in the US. For example, blue is attributed to males while pink is attributed to females [13, 15]. The predictions made by the model in recent years reflect this trend. It is also interesting to see how Pink in particular starts as masculine and becomes feminine around the 1950s as this is also when the gender stereotyping of the color changed. Pink was considered more of a boy's color and blue more of a girl's color until around the 1950s [13].

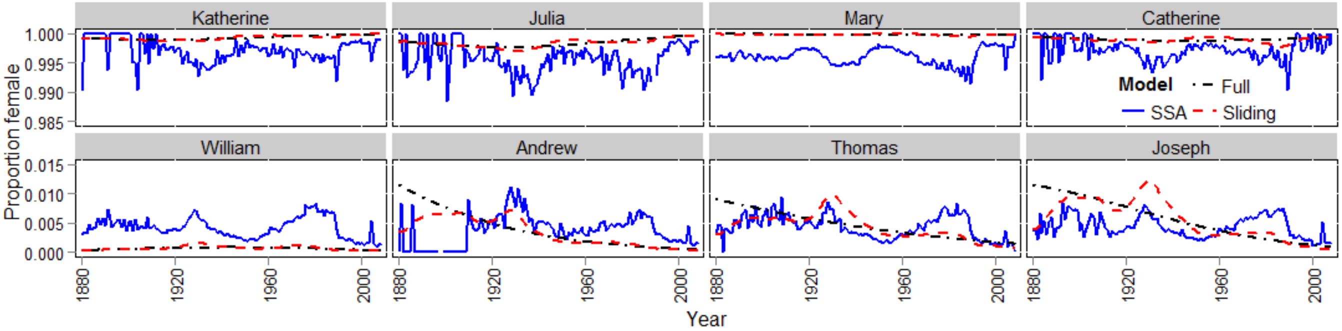

Figure 4. Predictions vs. SSA proportions for popular names given predominantly to either females or males over time.

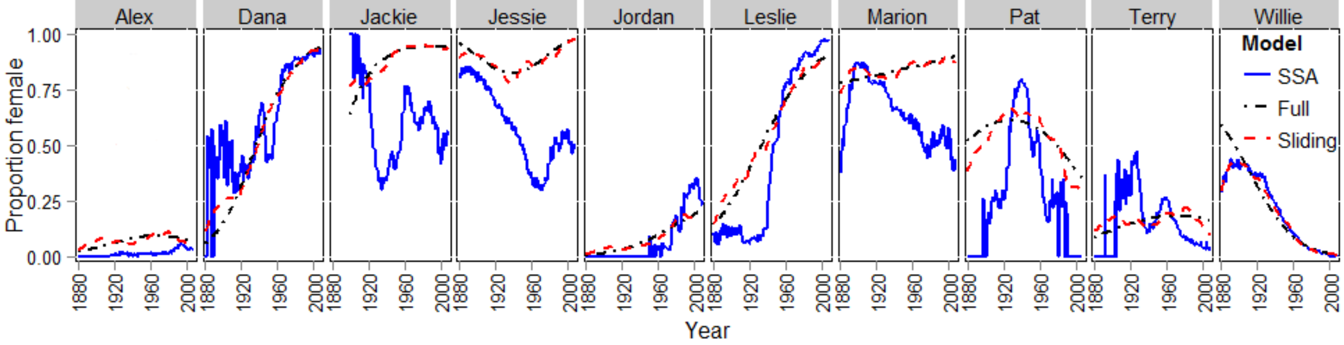

Figure 5. Predictions vs. SSA proportions for names given to both males and females at varying rates over time.

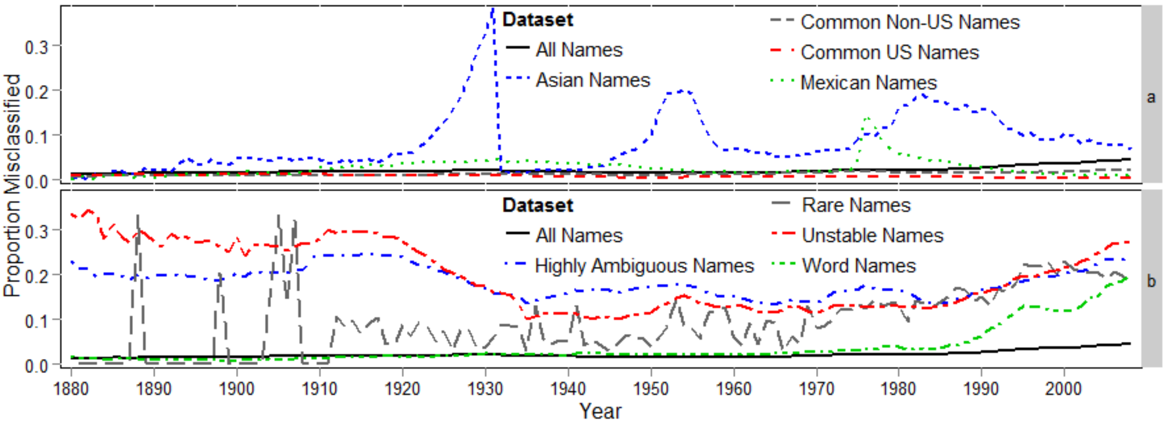

Figure 6. The proportion of people in the SSA dataset misclassified by the "full" model for different categories of names. a) *Asian names* [19]; *Common Non-US Names* correlated with non-US country affiliation in PubMed; *Common US names* with at least 500,000 individuals in the SSA dataset; *Mexican names* [17]. b) *Highly ambiguous names* have roughly equal males vs. females for any given year in the SSA dataset; *Rare names* given to < 11 people in the SSA dataset; *Unstable names* exclusively male at one point but changed to exclusively female (or vice versa); and *Word names* [32]. Some misclassification spikes in the Asian names and Mexican names reflect individual names: Jo (1931), Kim (1954), and Jaime (1976).

Two of the other colors, Brown and Green, are also popular surnames. As noted earlier, some of the prediction curves may be affected in unexpected ways simply because of the format of the search queries used to make the predictions. However, in this case, Brown and Green are names that one would expect to be more associated with males, and that is supported by the predictions generated by the model.

4.4.3 *Rare names*

The model performs worse with rare names (Figure 6b), here defined as names given to ten or fewer people in the SSA dataset. This is primarily because the Bing query counts tend to be low for all the phrase markers, although the low count indicator variables in the model help somewhat in these instances. Interestingly, many of the rare names are spelling variants of traditional, popular names (e.g., Ashlii, Ashliy, Ashlae, Ashlely vs. Ashley).

4.5 Robustness of model to potential changes in search engine coverage and estimation

A concern with using search engines is that their coverage, query processing, and count estimation change over time. To simulate how these changes might affect the model, we gathered Google search result counts for a sample of 2,645 names using the same search methods as with Bing. The set of names was sampled due to Google API limitations. One can think of this as a worst-case scenario simulation of Bing changes because the queries were submitted to Google at least a year before that of Bing data collection (i.e., reflects a temporal change of the web itself), and Google counts represent different web crawling methods, indexing, and count estimation. Figure 10 shows that using Google counts in the Bing optimized model introduces surprisingly little noise. Misclassifications go from ~1% to ~2% with some periodic variations. This suggests that the model is quite robust to changes in search engine coverage, query processing, and estimation over time.

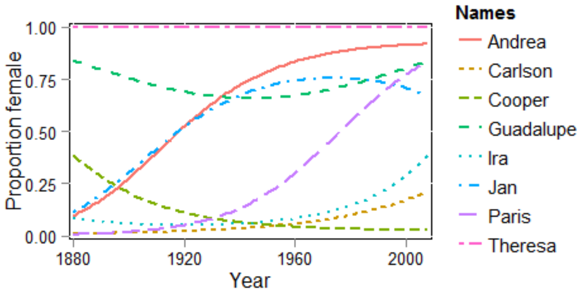

Figure 7. Predictions for some names with "atypical" patterns.

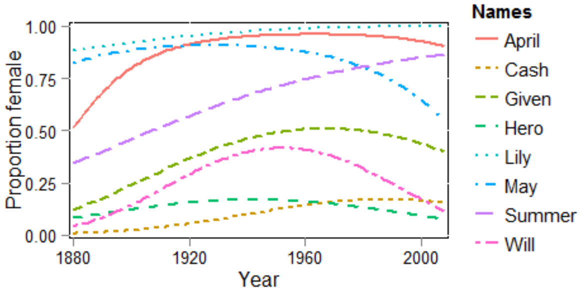

Figure 8. Predictions for names that are also common words.

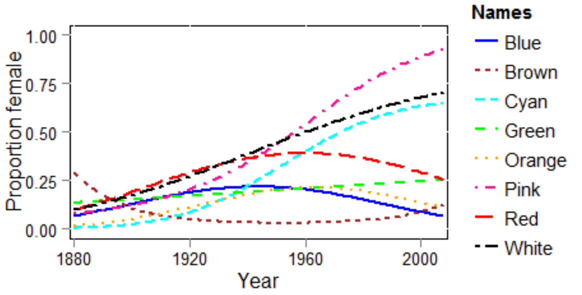

Figure 9. Predictions for names that are also colors.

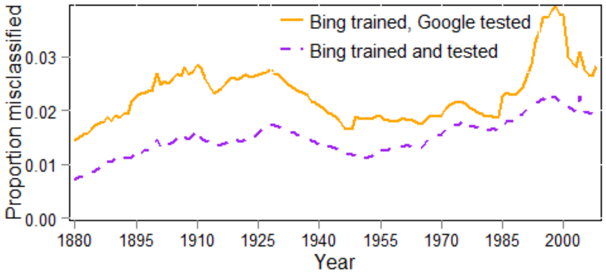

Figure 10. Potential degradation of model predictions due to changes in search engine coverage and count estimation as reflected by using Google counts in a Bing trained model.

4.6 Naming patterns over time

The statistical model allows one to study naming patterns over time so that periods of name popularity for individuals of specific genders and general trending patterns can be noted. For example, Figure 4, shows that Katherine, Julia, Mary, and Catherine have been strongly associated with females during the entire time period covered by the SSA dataset. However, Figure 5 shows that Dana and Leslie switched from being mostly male to mostly female. Several additional names in Figure 5 further illustrate the observation made by Lieberson, Dumais, and Baumann [16] that names considered androgynous are increasingly given to females over time. Willie exhibits the opposite trend. Pat, as well, seems to be becoming more associated with males. This may be because Pat is both a popular nickname for females (Patricia) and males (Patrick). As Patricia has decreased in popularity and Patrick has increased, the prediction curve for Pat is drawn towards that of Patrick and becomes more masculine over time. Alex is another example in this vein. Alex is a common nickname for both names associated with males, e.g., Alexander, and females, e.g., Alexandra, Alexa, and Alexandria. However, its formal use is most often associated with males. The model predict a majority of male but less so than the SSA data. In other words, even though the model was trained using formal given names (on SSA applications), we see that the model captures both the formal and informal (nickname) uses of each name (on the web).

4.7 Gender patterns over time in bibliographic databases

With a model in place we turn to illustrating its use in characterizing the temporal changes in gender for a selection of readily available bibliographic databases covering books, articles, patents, and grants: Harvard University Library, University of Illinois library, PubMed (biomedical articles), DBLP (computer science articles), patents from the United States Patent and Trademark Office (USPTO), and grants from the US Department of Health and Human Services (HHS), primarily NIH. For each of the bibliographic datasets we only considered records with a publication date from 1800 to 2010 that had at least one author (or inventor, or principal investigator) name and only assigned the gender to the first listed author (or inventor, or investigator). Gender was assigned if the model produced high confidence classification and these classification were consistent over time (female if $p > 0.9$, male if $p < 0.1$ for all of the years 1955, 1965, 1975; unknown otherwise), and the name was already in the Genni database. Each dataset contains millions of records: Harvard Library Bibliographic Dataset [14]: ~8 million records, 1800-2010; University of Illinois Library (provided by library staff): ~3 million records, 1800-2009; PubMed (leased for free from the NLM): 18+ million records, 1865-2010. Note that PubMed started consistently recording first names only in 2002, so we used a disambiguated dataset of authors [24, 25] to assign a first name to older records that were matched with complete records. DBLP [10]: ~1.7 million records from major computer science journals and conference proceedings, 1936-2010. NIH Exporter CRISP legacy grant data [8]: ~2 million records, 1970-2009. USPTO patents distribution by Google [28]: ~4 million records, 1976-2010.

Figure 11 shows two plots reflecting trends over time. The top plot shows the proportion of records that were assigned with high confidence within each dataset, and the bottom plot shows the proportion of high confidence classifications that were female. Note that the spikes in these curves often reflect low counts. For example, the 1878 spike in PubMed (and the 1940 spike in DBLP) in the bottom plot represents only 5 people.

Two major driving forces can explain the patterns for these curves: indexing practices and human behavior, both of which may exhibit gender bias. We expect the top plot to primarily reflect indexing practice, and the bottom plot to primarily reflect human behavior, more specifically participation bias. For example, NIH and USPTO have consistently indexed nearly all of the records with first names over time, while DBLP and the two libraries have been consistent over time but at a lower rate. PubMed first started recording first names in 2002, and even though disambiguation helps assign gender to some older records, we see a dramatic increase over time. The recent downward trend in DBLP, USPTO, and PubMed is not due to indexing practice but rather an increasing presence of non-US names for which the model predicts neutral. In analyzing these gender trends, we are confident in the statistics for the datasets that are "complete" in the sense that USPTO covers all US patents, and CRISP covers all extramural NIH grants. The other datasets reflect selective inclusion driven and constrained by a variety of factors. For example, PubMed (and DBLP) probably aim at covering the most important publication venues in biomedicine (and computer science, respectively), whereas the libraries cater to their (local) patrons. All are constrained by limited resources. One might wonder how much of the selective inclusion reflects gender bias, beyond the general participation bias in these specific domains (biomedicine and computer science).

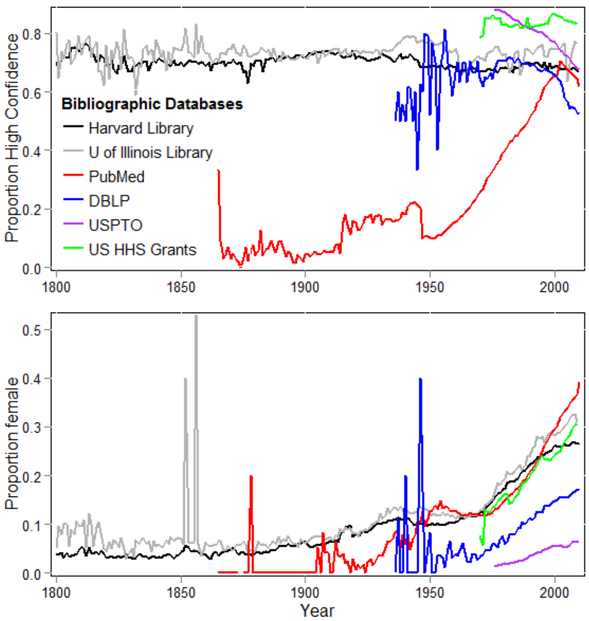

Figure 11. Gender classifications in bibliographic databases.

In terms of participation, all domains reflect a bias towards male. USPTO clearly has the lowest female participation (7% peak), followed by DBLP (17% peak), Harvard Library (27% peak), NIH grants (31% peak), Illinois Library (31% peak), and PubMed (39% peak). Female participation has consistently increased in all datasets since 1970, although the growth varies: USPTO has slow growth, DBLP moderate growth, and the remaining datasets have fast growth. If the trends continue, PubMed will reach the magical 50% female in ~10-20 years, DBLP in ~100 years, and USPTO in ~300 years. These are obviously very rough estimates that rest on a suite of assumptions that probably will not hold in the future. For example, a) Harvard library already appears to have stagnated about 10 years ago; b) females on NIH grants and PubMed followed each other closely from the 1970s to the mid-1990s when NIH suffered a setback and has been trying to catch up ever since. These global patterns hint at potential in-depth studies addressing confounding factors. For example, among younger authors in PubMed, today's proportion female is probably closer to 50%.

5. CONCLUSIONS AND FUTURE WORK

Discerning the gender orientation of a first name given no other information can be a difficult task. Our approach performs surprisingly well despite its simplicity. It misclassifies 2.25% of the people in the SSA dataset, which is just slightly higher than the 1.74% pure error rate, and it utilizes simple web searches that are highly scalable and its predictions are based on public data covering names beyond the SSA dataset.

Very rare names are the most challenging for our approach because of limited search engine coverage. Performance would improve if the search engine could: 1) provide higher recall by covering more of the surface web, if not the deep web, 2) provide higher precision by permitting exact phrase searching (to avoid cases where punctuation occurs between words in a phrase), and 3) provide better estimates of result counts. Other groups of names have high rates of misclassification as well. For names that are intrinsically ambiguous (or made so by many-to-many transliteration from the original language into English), no model can do better without contextual clues surrounding the mention of the name. Other names like Andrea and Ira are globally androgynous but not so locally because of different naming conventions across countries and cultures. A natural extension of our work is to assess the promise of additional markers (for gender, time, geography/language) and relaxing the strict phrase searching for predicting the culture or geographic origin of names, either alone or in conjunction with gender. One could also combine our approach with name features extracted directly from the name and use other sources for training data such as Wikipedia or Facebook (as demonstrated by Treeratpituk and Giles [26] and Tang et al. [23]). Other sources of training data and new attributes could also help improve predictions for diminutive names or nicknames such as Alex which serves as a nickname for a female while it is also a formal given name for a male. Future work could also study more in-depth some of the (unexpected) temporal patterns revealed by our model.

Nevertheless, the present model permits accurate and automatic assignment of gender to names beyond the public SSA dataset. The model has been implemented as a web-tool called Genni (available via http://abel.lis.illinois.edu/) that displays the predicted proportion of females vs. males over time for any given name. This should be a valuable resource for those who utilize names in order to discern gender on a large scale, e.g., to study the roles of gender and diversity in scholarly work based on digital libraries and bibliographic databases where the authors' names are listed.

6. ACKNOWLEDGMENTS

Funding for this study was provided in part by NSF Science of Science and Innovation Policy (SciSIP) program award 0965341.

7. REFERENCES

[1] API Basics: 2011. *http://www.bing.com/developers/s/APIBasics.html*. Accessed: 2012-05-21.

[2] Baby Name Guesser - analysis of first names, first names statistics, popular boy and girl names, male names, female names: *http://www.gpeters.com/names/baby-names.php*. Accessed: 2011-05-03.

[3] Baby Names: 2011. *http://www.babynamewizard.com/*. Accessed: 2011-05-03.

[4] Bergsma, S. et al. 2009. Glen, Glenda or Glendale: unsupervised and semi-supervised learning of English noun gender. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (Boulder, Colorado, 2009), 120–128.

[5] Bing API, Version 2: *http://msdn.microsoft.com/en-us/library/dd251056.aspx*. Accessed: 2013-04-04.

[6] Breuning, M. and Sanders, K. 2007. Gender and Journal Authorship in Eight Prestigious Political Science Journals. *PS: Political Science & Politics*. 40, 02 (Apr. 2007), 347.

[7] Carlson, A. et al. 2010. Toward an Architecture for Never-Ending Language Learning. *Proceedings of the AAAI-10 24th AAAI Conference on Artificial Intelligence* (2010).

[8] CRISP LEGACY DATA: 2010. *http://exporter.nih.gov/crisp_catalog.aspx*. Accessed: 2013-04-10.

[9] Cucerzan, S. and Yarowsky, D. 2003. Minimally supervised induction of grammatical gender. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* (Edmonton, Canada, 2003), 40–47.

[10] DBLP FAQ: What is the meaning of "DBLP"?: 2013. *http://www.informatik.uni-trier.de/~ley/db/about/faqdblp.html*. Accessed: 2013-04-04.

[11] Efron, M. 2004. The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. *Proceedings of the 13th ACM international conference on Information and knowledge management* (Washington D.C., USA, 2004), 390–398.

[12] England, P. et al. 2007. Why Are Some Academic Fields Tipping Toward Female? The Sex Composition of U.S. Fields of Doctoral Degree Receipt, 1971–2002. *Sociology of Education*. 80, 1 (2007), 23–42.

[13] Frassanito, P. and Pettorini, B. 2008. Pink and blue: the color of gender. *Child's Nervous System*. 24, 8 (2008), 881–882.

[14] Harvard Library Bibliographic Dataset | Open Metadata: 2012. *http://openmetadata.lib.harvard.edu/bibdata*. Accessed: 2013-04-04.

[15] Karniol, R. 2011. The Color of Children's Gender Stereotypes. *Sex Roles*. 65, 1 (2011), 119–132.

[16] Lieberson, S. et al. 2000. The Instability of Androgynous Names: The Symbolic Maintenance of Gender Boundaries. *The American Journal of Sociology*. 105, 5 (Mar. 2000), 1249–1287.

[17] Mexican Babies Name List: *http://www.top-100-baby-names-search.com/mexican-babies-name.html*. Accessed: 2013-04-04.

[18] Otterbacher, J. 2010. Inferring gender of movie reviewers: exploiting writing style, content and metadata. *Proceedings of the 19th ACM international conference on Information and knowledge management* (Toronto, ON, Canada, 2010), 369–378.

[19] Popular Asian Names: *http://www.top-100-baby-names-search.com/popular-asian-names.html*. Accessed: 2013-04-04.

[20] Popular Baby Names: 2010. *http://www.ssa.gov/oact/babynames/*. Accessed: 2011-02-17.

[21] Reece-Evans, L. 2010. Gender and Citation in Two LIS E-Journals: A Bibliometric Analysis of LIBRES and Information Research. *Library & Information Science Research Electronic Journal*. 20, 1 (Mar. 2010), 1–18.

[22] Sassen, C. 2009. Gender and authorship in The Indexer, 1958-2007. *Indexer*.

[23] Tang, C. et al. 2011. What's in a name: a study of names, gender inference, and gender behavior in facebook. *Proceedings of the 16th international conference on Database systems for advanced applications* (Hong Kong, China, 2011), 344–356.

[24] Torvik, V.I. et al. 2005. A probabilistic similarity metric for Medline records: A model for author name disambiguation: Research Articles. *Journal of the American Society for Information Science and Technology*. 56, 2 (Jan. 2005), 140–158.

[25] Torvik, V.I. and Smalheiser, N.R. 2009. Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 3, 3 (Jul. 2009), 11:1–11:29.

[26] Treeratpituk, P. and Giles, C.L. 2012. Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching. *Proceedings of the 26th AAAI Conference on Artificial Intelligence* (2012), 1141–1147.

[27] Turney, P.D. and Littman, M.L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.* 21, 4 (2003), 315–346.

[28] USPTO Bulk Downloads: Patent Grant Bibliographic Data: *http://www.google.com/googlebooks/uspto-patents-grants-biblio.html*. Accessed: 2013-04-29.

[29] Vogel, A. and Jurafsky, D. 2012. "He Said, She Said: Gender in the ACL Anthology. *ACL 2012 Special Workshop: Rediscovering 50 Years of Discoveries* (2012).

[30] Wattenberg, M. 2005. Baby Names, Visualization, and Social Data Analysis. *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization* (2005), 1.

[31] Web 1T 5-gram Version 1: 2006. *http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13*. Accessed: 2013-04-04.

[32] Word Names: which words are being used for Brit Babies?: 2011. *http://britishbabynames.typepad.com/blog/2011/05/word-names-which-words-are-being-used-for-brit-babies.html*. Accessed: 2013-04-04.

[33] Zheleva, E. and Getoor, L. 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. *Proceedings of the 18th international conference on World wide web* (Madrid, Spain, 2009), 531–540.